

Содержание:

image not found or type unknown



ВВЕДЕНИЕ

Целью курсовой работы является изучение технологии OLAP, понятие ее реализации и структуры.

В современном мире компьютерные сети и вычислительные системы позволяют анализировать и обрабатывать большие массивы данных.

Большой объем информации сильно усложняет поиск решений, но дает возможность получить намного точнее расчеты и анализ. Для решения такой проблемы существует целый класс информационных систем, выполняющих анализ. Такие системы называют системами поддержки принятия решений (СППР) (DSS, Decision Support System).

Для выполнения анализа СППР должна накапливать информацию, обладая средствами ее ввода и хранения. Всего можно выделить три основные задачи, решаемые в СППР:

ввод данных;

хранение данных;

анализ данных.

Ввод данных в СППР осуществляется автоматически от датчиков, характеризующих состояние среды или процесса, или человеком-оператором.

Если ввод данных осуществляется автоматически от датчиков, то данные накапливаются по сигналу готовности, возникающему при появлении информации или путем циклического опроса. Если же ввод осуществляется человеком, то они должны предоставлять пользователям удобные средства для ввода данных, проверяющих их на правильность ввода, а так же выполнять необходимые вычисления.

При вводе данных одновременно несколькими операторами, необходимо решать проблемы модификации и параллельного доступа одних и тех же данных.

СППР предоставляет аналитику данные в виде отчетов, таблиц, графиков для изучения и анализа, именно поэтому такие системы обеспечивают выполнение функции поддержки принятия решений.

В подсистемах ввода данных, называемых OLTP (On-line transaction processing), реализуется операционная обработка данных. Для их реализации используют обычные системы управления БД (СУБД).

Подсистема анализа может быть построена на основе:

подсистемы информационно-поискового анализа на базе реляционных СУБД и статических запросов с использованием языка SQL;

подсистемы оперативного анализа. Для реализации таких подсистем применяется технология оперативной аналитической обработки данных OLAP, использующая концепцию многомерного представления данных;

подсистемы интеллектуального анализа. Данная подсистема реализует методы и алгоритмы Data Mining¹.

С точки зрения пользователя, OLAP-системы представляют средства гибкого просмотра информации в различных срезах, автоматического получения агрегированных данных, выполнения аналитических операций свёртки, детализации, сравнения во времени. Благодаря всему этому OLAP-системы являются решением с большими преимуществами в области подготовки данных для всех видов бизнес-отчетности, предполагающих представление данных в различных разрезах и разных уровнях иерархии, таких как, отчетов по продажам, различных форм бюджетов и других. OLAP-системы имеет большие плюсы подобного представления и в других формах анализа данных, в том числе для прогнозирования.

1. OLAP-СИСТЕМЫ

1.1. Многомерная модель данных

Подсистемами анализа решают проблемы связанные с организацией эффективного анализа данных и предоставления к ним доступа. В процессе принятия решений пользователь составляет некоторые гипотезы чтобы они стали законченными решениями необходимо провести их. Проверка гипотез осуществляется на основании информации об анализируемой предметной области. Обычно наиболее удобный способ представления такой информации для человека является зависимость между некоторыми параметрами.

В процессе анализа данных, поиска решений часто появляется необходимость в построении зависимостей между различными параметрами, но так как число таких параметров может изменяться в широких пределах, а традиционные средства анализа, работающие с данными, представленными в виде таблиц реляционной БД, не могут полностью соответствовать этим требованиям. В 1993 г. Э. Ф. Кодд — основоположник реляционной модели БД — рассмотрел ее недостатки в первую очередь он указал на невозможность "объединять, просматривать и анализировать данные с точки зрения множественности измерений, т. е. самым понятным для аналитиков способом". Измерение — это последовательность значений одного из анализируемых параметров. Множественность измерений предполагает представление данных в виде многомерной модели. По измерениям в многомерной модели откладывают параметры, относящиеся к анализируемой предметной области.

По Кодду, многомерное концептуальное представление есть множественная перспектива, состоящая из нескольких независимых измерений, вдоль которых могут быть проанализированы определенные совокупности данных.

Одновременный анализ по нескольким измерениям определяется как многомерный анализ¹, где измерение можно представлять не только в виде иерархической структуры, но и в некоторых случаях как несколько видов иерархического представления.

На пересечениях осей измерений располагаются данные, количественно характеризующие анализируемые факты, — меры, такие как издержки, остатки на складе или же объемы продаж, выраженные в единицах продукции или в денежном выражении и т. п.

Многомерную модель данных можно представить как гиперкуб (приложение 1, а) (конечно, название не очень удачно, поскольку под кубом обычно понимают фигуру с равными ребрами, что в данном случае далеко не так). Ячейками гиперкуба,

представляющего модель данных являются меры, а ребрами – измерения. Благодаря такой модели данных пользователи могут получать подмножества данных, формулировать сложные запросы, генерировать отчеты.

Над гиперкубом могут выполняться три типа операций:

Срез (приложение 1, b) — формируется подмножество многомерного массива данных, которое должно соответствовать единственному значению нескольких или даже одного элементов измерений, не входящих в это подмножество. Если рассматривать термин «Срез» со стороны конечного пользователя, то чаще всего его роль выполняет двумерная проекция куба.

Вращение (приложение 2) — изменение расположения измерений, представленных на отображаемой странице или же в отчете. Кроме того, вращением куба данных является перемещение внетабличных измерений на место измерений, представленных на отображаемой странице, и наоборот (при этом внетабличное измерение становится новым измерением строки или измерением столбца). В качестве примера для первого случая может служить такой отчет, для которого элементы измерения "Время" располагаются поперек экрана (являются заголовками столбцов таблицы), а элементы измерения "Производство" — вдоль экрана (являются заголовками строк таблицы). После применения операции вращения отчет будет иметь следующий вид: элементы измерения "Производство" будут расположены по горизонтали, а элементы измерения "Время" — по вертикали. Примером второго случая может служить преобразование отчета с измерениями "Меры" и "Производство", расположенными по вертикали, и измерением "Время", расположенным по горизонтали, в отчет, у которого измерение "Меры" располагается по вертикали, а измерения "Время" и "Производство" — по горизонтали. При этом элементы измерения "Время" располагаются над элементами измерения "Производство". Для третьего случая применения операции вращения можно привести пример преобразования отчета с расположенным по горизонтали измерением "Время" и измерением "Производство", расположенным по вертикали, в отчет, у которого по горизонтали представлено измерение "Время", а по вертикали — измерение "География" (синоним: Pivot).

Консолидация и детализация (приложение 3) — операции, определяющие переход вверх по направлению от детального (down) представления данных к агрегированному (up) и наоборот, соответственно. Направление детализации (обобщения) может быть задано как по иерархии отдельных измерений, так и согласно прочим отношениям, установленным в рамках измерений или между

измерениями. Например, если при анализе данных об объемах продаж в Северной Америке выполнить операцию Drill Down для измерения "Регион", то на экране будут отображены такие его элементы, как "Канада", "Восточные Штаты Америки" и "Западные Штаты Америки". В результате дальнейшей детализации элемента "Канада" будут отображены элементы "Торонто", "Ванкувер", "Монреаль" и т. д1.

1.2 Определение OLAP-систем

Технология комплексного многомерного анализа данных получила название OLAP. OLAP — это ключевой компонент организации ХД.

OLAP-функциональность может быть реализована различными способами, как простейшими, такими как анализ данных в офисных приложениях, так и более сложными - распределенными аналитическими системами, основанными на серверных продуктах.

OLAP (On-Line Analytical Processing) - технология оперативной аналитической обработки данных использующая средства и методы для сбора, хранения и анализа многомерных данных и целях поддержки процессов принятия решений.

Основное назначение OLAP-систем — поддержка аналитической деятельности, произвольных запросов пользователей-аналитиков. Целью OLAP-анализа является проверка возникающих гипотез.

В 1993 году Эдгаром Коддом опубликовал статью «OLAP для пользователей-аналитиков: каким он должен быть». В работе Кодд изложил основные концепции оперативной аналитической обработки и определил двенадцать правил, которым должны удовлетворять продукты, позволяющие выполнять оперативную аналитическую обработку:

Многомерность;

Прозрачность;

Доступность;

Постоянная производительность при разработке отчетов;

Клиент-серверная архитектура;

Равноправие измерений;

Динамическое управление разреженными матрицами;

Поддержка многопользовательского режима;

Неограниченные перекрестные операции;

Интуитивная манипуляция данными;

Гибкие возможности получения отчетов;

Неограниченная размерность и число уровней агрегации.

2. КОНЦЕПТУАЛЬНОЕ МНОГОМЕРНОЕ ПРЕДСТАВЛЕНИЕ

2.1. Двенадцать правил Кодда

1. Многомерность — OLAP-система на концептуальном уровне должна представлять данные в виде многомерной модели, так как это сильно упрощает процессы восприятия и анализа информации.

2. Прозрачность — OLAP-система должна скрывать от пользователя реальную реализацию многомерной модели, способ организации, источники, средства обработки и хранения.

3. Доступность — OLAP-система должна предоставлять пользователю единую, согласованную и целостную модель данных, обеспечивая доступ к данным независимо от места и способа их хранения.

4. Постоянная производительность при разработке отчетов — производительность OLAP-систем не должна значительно уменьшаться при увеличении количества измерений, по которым выполняется анализ.

5. Клиент-серверная архитектура — OLAP-система должна быть способна работать в среде "клиент-сервер", т. к. большинство данных, которые требуется подвергать оперативной аналитической обработке, хранятся распределено. Главной идеей является то, что серверный компонент инструмента OLAP должен быть достаточно

интеллектуальным и позволять строить общую концептуальную схему на основе консолидации и обобщения различных физических и логических схем корпоративных БД для обеспечения эффекта прозрачности.

6. **Равноправие измерений** — OLAP-система должна поддерживать многомерную модель, в которой все измерения равноправны. При необходимости дополнительные характеристики могут быть предоставлены отдельным измерениям, но такая возможность должна быть предоставлена любому измерению.

7. **Динамическое управление разреженными матрицами** — OLAP-система должна обеспечивать оптимальную обработку разреженных матриц. Скорость доступа должна сохраняться вне зависимости от расположения ячеек данных и быть постоянной величиной для моделей, имеющих разное число измерений и различную степень разреженности данных.

8. **Поддержка многопользовательского режима** — OLAP-система должна предоставлять возможность работать нескольким пользователям совместно с одной аналитической моделью или создавать для них различные модели из единых данных. Из-за возможности чтения и записи данных, система должна обеспечивать целостность и безопасность информации.

9. **Неограниченные перекрестные операции** — OLAP-система должна обеспечивать сохранение функциональных отношений, описанных с помощью определенного формального языка между ячейками гиперкуба при выполнении любых из возможных четырех операций: среза, вращения, консолидации или детализации. Система должна самостоятельно выполнять преобразование установленных отношений, не требуя от пользователя их переопределения.

10. **Интуитивная манипуляция данными** — OLAP-система должна предоставлять способ выполнения операций среза, вращения, консолидации и детализации над гиперкубом без необходимости пользователю совершать множество действий с интерфейсом. Измерения, определенные в аналитической модели, должны содержать всю необходимую информацию для выполнения всех возможных операций над гиперкубом.

11. **Гибкие возможности получения отчетов** — OLAP-система должна поддерживать различные способы визуализации данных, т. е. отчеты должны представляться в любой возможной ориентации. Средства формирования отчетов должны представлять синтезируемые данные или информацию, следующую из модели

данных в ее любой возможной ориентации.

12. Неограниченная размерность и число уровней агрегации — исследование о возможном числе необходимых измерений, требующихся в аналитической модели, показало, что одновременно может использоваться до девятнадцать измерений. Отсюда вытекает настоятельная рекомендация, чтобы аналитический инструмент мог одновременно предоставить не меньше пятнадцати измерений, а предпочтительнее двадцати измерений. Более того, каждое из общих измерений не должно быть ограничено по числу определяемых пользователем-аналитиком уровней агрегации и путей консолидации.

3. АРХИТЕКТУРА OLAP-СИСТЕМ

Многомерность в OLAP-приложениях представляют в виде трех уровней:

Многомерное представление данных - средства конечного пользователя, обеспечивающие многомерную визуализацию и манипулирование данными; слой многомерного представления абстрагирован от физической структуры данных и воспринимает данные как многомерные.

Многомерная обработка - средство (язык) формулирования многомерных запросов (традиционный реляционный язык SQL здесь оказывается непригодным) и процессор, умеющий обработать и выполнить такой запрос.

Многомерное хранение - средства физической организации данных, обеспечивающие эффективное выполнение многомерных запросов.

Во всех OLAP-системах первые два уровня присутствуют обязательно, а третий уровень не обязательно присутствует в них, хотя и является широко распространенным, так как данные для многомерного представления могут извлекаться и из обычных реляционных структур, и тогда процессор многомерных запросов будет транслировать многомерные запросы в SQL-запросы, которые выполняются реляционной СУБД.

OLAP-продукты, чаще всего, представляют собой OLAP-сервер, многомерную серверную СУБД (такую как, Microsoft OLAP Services или Oracle Express Server) или же OLAP-клиент, средство многомерного представления данных (такое как, Pivot Tables в Excel 2000 фирмы Microsoft или ProClarity фирмы Knosys)¹.

OLAP-сервер обеспечивает хранение данных, выполнение над ними необходимых операций и формирование многомерной модели на концептуальном уровне.

OLAP-клиент обеспечивает пользователю возможность удобно манипулировать данными для выполнения задач анализа, представляя пользователю интерфейс к многомерной модели данных².

Слой многомерной обработки обычно бывает встроен в OLAP-клиент и/или в OLAP-сервер, но так же этот слой может быть выделен в чистом виде, как, например, компонент Pivot Table Service фирмы Microsoft.

OLAP-серверы, или серверы многомерных БД, могут хранить свои многомерные данные по-разному. В любом ХД как в обычном так и в многомерном вместе с детальными данными, извлекаемыми из оперативных систем, хранятся и суммарные показатели (агрегированные показатели, агрегаты), такие, как суммы объемов продаж по месяцам, по категориям товаров и т. п. Агрегаты хранятся в явном виде с единственной целью - ускорить выполнение запросов.

Как детальные данные, так и агрегаты могут храниться либо в реляционных, либо в многомерных структурах. Многомерное хранение позволяет обращаться с данными как с многомерным массивом, благодаря чему обеспечиваются одинаково быстрые вычисления суммарных показателей и различные многомерные преобразования по любому из измерений. Некоторое время назад OLAP-продукты поддерживали либо реляционное, либо многомерное хранение. Сегодня, как правило, один и тот же продукт обеспечивает оба этих вида хранения, а также третий вид - смешанный. Способ реализации хранения данных очень важен, т. к. от него зависят такие характеристики, как занимаемые ресурсы и в следствии производительность. Выделяют три основных способа реализации:

MOLAP (Multidimensional OLAP) - детальные данные и агрегаты хранятся в многомерной БД. В этом случае получается наибольшая избыточность, так как многомерные данные полностью содержат реляционные.

ROLAP (Relational OLAP) - детальные данные остаются в реляционной БД; агрегаты хранятся в той же БД в специально созданных служебных таблицах.

HOLAP (Hybrid OLAP) - детальные данные остаются в реляционной БД, а агрегаты хранятся в многомерной БД.

Каждый из этих способов имеет свои преимущества и недостатки и должен применяться в зависимости от различных условий - объема данных, мощности реляционной СУБД и т. д.

При хранении данных в многомерных структурах возникает потенциальная проблема «разбухания» за счет хранения пустых значений. Если в многомерном массиве зарезервировано место под все возможные комбинации меток измерений, а реально заполнена лишь малая часть, то большая часть куба будет пустовать, хотя место будет занято. Современные OLAP-продукты умеют справляться с этой проблемой¹.

ЗАКЛЮЧЕНИЕ

Таким образом, в ходе рассматриваемой темы мы узнали, что системами поддержки принятия решений может быть построена на подсистеме оперативного анализа. Для реализации которой и применяется технология оперативной аналитической обработки данных OLAP (On-line analytical processing), использующая концепцию многомерного представления данных.

Узнали, что множественность измерений предполагает представление данных в виде многомерной модели. Измерение — это последовательность значений одного из анализируемых параметров. По измерениям в многомерной модели откладывают параметры, относящиеся к анализируемой предметной области.

По Кодду, многомерное концептуальное представление — это множественная перспектива, состоящая из нескольких независимых измерений, вдоль которых могут быть проанализированы определенные совокупности данных.

Одновременный анализ по нескольким измерениям определяется как многомерный анализ.

Оказалось, что многомерную модель данных можно представить как гиперкуб. Ребрами которого являются измерения, а ячейками — меры.